

Job Detection in Twitter

Besat Kassaie

Spring 2016

Abstract

In this report, we propose a new application for twitter data called *job detection*. We identify people's job category based on their tweets. As a preliminary work, we limited our task to identify only IT workers from other job holders. We have used and compared both simple bag of words model and a document representation based on Skip-gram model. Our results show that the model based on Skip-gram, achieves a 76% precision and 82% recall.¹

1 Introduction

Internet users are producing large amount of data with almost no cost for people and companies who can exploit this data for their own benefits. Besides, more sophisticated data analysis techniques are available nowadays comparing to even one decade ago. Using data analysis techniques invaluable information can be induced from abundant data available on the web.

Social networks, such as twitter, are one of the most popular class of applications gathering a lot of information in different formats such as text, image, and video. So far people have worked on twitter data from interesting and different aspects. They could extract highly accurate information in terms of sentiments [5], fine grain location information [6], churn prediction [2], topic detection [8], and so on.

Our main contribution in this work is to detect twitter users' job based on the textual content of their tweets. We did not find any similar work to target this application on twitter. Job detection from tweets have many potential applications such as targeted commercial advertisement, credit scoring, and so on.

We faced some challenges in this project. First, there is a huge diversity among career domains, from cosmetic services to medical fields and from astronauts to miners. We would need a huge set of samples from twitter users to cover all of these categories. On the other hand, it is not feasible to crawl such dataset in a short time, considering the limitations which twitter imposes on the rate of fetching tweets and also our hardware resources. The other challenge is

¹This document is the project report prepared for CS 886, University of Waterloo

that we need to label each user with a job category in our dataset. This is not easy task to produce such a training data to cover all job categories. Also, not all people reveal their information in twitter due to their career category. For example it could be assumed that journalists are more likely to have an active twitter account than miners.

Considering all those challenges, we limited the target job categories into Information Technology related jobs. By this assumption, we need to gather much less data than what is needed to cover all job categories. Also we can assume that many of IT workers are likely to use twitter. Finally as we are familiar with job titles in this field, we could label data easily and accurately.

Another contribution of this work is applying Skip-gram model for computing word vectors and using KMeans for representing documents by word vectors. We showed that our model based on word vectors outperforms simple bag of word representation of documents.

In the next sections we first introduce our data gathering strategies and methods. Then we present the implementation details and results and finally we give the conclusions and future works.

2 Method and Data

For this work we needed a set of twitter users labeled with their job. As there is no such dataset we had to create our own dataset. Compiling such dataset for all jobs would take a lot of time and resources. So in this preliminary work we focused on detecting people with IT jobs. So our labeled dataset would include people labeled as IT workers or non IT workers.

In this work we investigate two architectures for classification based on different approaches for document representation. The first one relies on the well-known bag of words model for document representation. For the second model, we use a document representation based on the term vectors which are extracted by word2vec. Word2vec is a tool for computing vector representations of words introduced by a team of researchers at Google. We explain more about word2vec in the next section.

We also propose more in detail explanation for data gathering strategies as well as our document representation and classification techniques in next sections.

2.1 Word2Vec

Although representing words as indices in the dataset vocabulary for using in NLP tasks has many advantages such as simplicity and fast model creation, it ignores possible and obvious similarities between words. For example the simple techniques of word representation cannot detect the semantic similarity between ‘King’ and ‘Man’ as well as the syntactic similarities between ‘Flowers’ and ‘Cats’. The ideal word representation for many applications is a representation which is able to detect all possible similarities and also preserve regularities

between vectors as much as possible. The regularities are observed as constant vector offsets between pairs of words sharing a particular relationship [7]. Some examples of these regularities are listed below:

$$\begin{aligned} \text{vector ("King")} - \text{vector ("Man")} + \text{vector ("Woman")} &= \text{vector ("Queen")} \\ \text{vector ("apple")} - \text{vector ("apples")} &= \text{vector ("car")} - \text{vector ("cars")} \end{aligned}$$

In order to capture these regularities there are different models such as Bag-of-Words Model (known as continuous bag of words, or CBOW) and Skip-gram Model (Figure 2.1) . The first model uses context to predict a target word and the second model uses a word to predict a target context.

We use the Skip-gram method because it produces more accurate results on large datasets. In this paper we used Word2Vec [1] as the tool for computing vector representations of words which uses the Skip-gram model.

In the Skip-gram model each current word is used as an input to a log-linear classifier with continuous projection layer to predict words within a certain range before and after the current word. Apparently a larger range results in better word vectors and at the same time imposes more computational costs. As the distance of context words from the current word is increased they get less related to the current word so Skip-gram model assigns less weight to the distant words. To do so the Skip-gram model samples less from distant words in the training dataset.

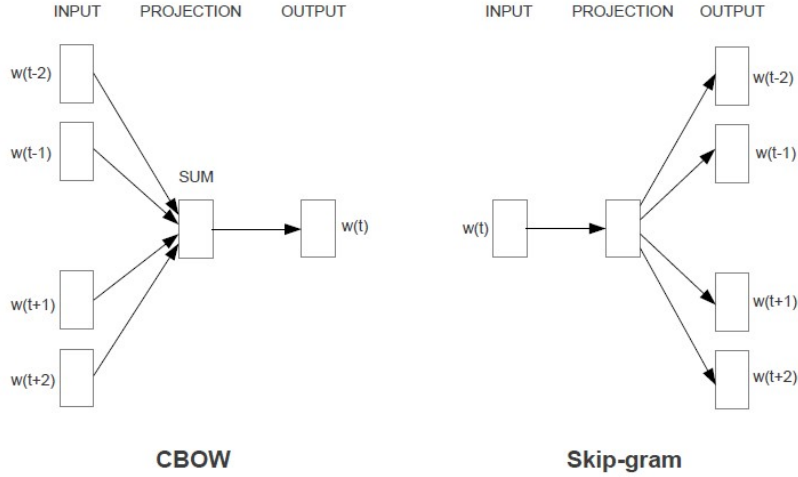


Figure 1: The CBOW architecture predicts the current word based on the context, and the Skip-gram predicts surrounding words given the current word [7]

The Skip-gram model trains high dimensional word vectors on a large dataset and detects very accurate semantic relationships between words which can be

applied on NLP applications and results in surprising results.

2.2 Data Preparation

We used two techniques for building our dataset. As our first approach we tried to use LinkedIn profiles for obtaining some auto labeled users. In this approach we look for people in LinkedIn who indicated that they are working in some IT related jobs. Then we could use their name and try to find an associated twitter ID to their LinkedIn name.

There are some challenges in this approach. First of all there are no dictionary of IT job titles. To deal with this issue we compiled a set of job names to cover such jobs. This dictionary includes 183 job titles. Next challenge is that there is no free API in LinkedIn for searching over people. To deal with this challenge we used the free Bing search engine API. We used a search query like: $\{jobtitle\} + 'site:ca.linkedin.com/in'$ for obtaining name of people in LinkedIn who mentioned one of our job titles in the description of their profile. We gathered about 4092 of such LinkedIn users for our IT job titles. The final challenge is to match this names with twitter IDs. For this part we first used UserSearch API in twitter. Using that API we gathered 43719 candidate twitter IDs. As there are multiple twitter IDs for each LinkedIn ID we need to filter the twitter IDs. To do so, in the next step, we got the profile data of the candidate twitter IDs. We used the description of these profiles and calculate their *Jaccard similarity* scores to the description of LinkedIn profiles. We filtered out the twitter IDs which had a similarity score under 0.5. By this method we gathered a total of 277 twitter IDs with IT related jobs. We gathered the most recent 3200 tweets of this users and included them in our dataset.

The second approach is a manual method for gathering data in which we used the twitter API directly. Here we again used the UserSearch API in twitter and queried using each job title. The gathered data were not clean and many of the twitter IDs were not actually related to people working in IT jobs. We used a manual process of verifying each of this candidate twitter IDs based on the description of their twitter profiles. By using both previous approaches we gathered a total of 805 positive and 574 negative samples. Like the previous method we fetched the most recent 3200 tweets per user for both set of users.

In our work we also needed some unlabeled data for a pretraining phase. We gathered the last 3200 tweets of for each of about 7237 random twitter users for this purpose.

2.3 Classification Architectures

In our work we do not look for the signals in a single tweet level, we combine the user's most recent tweets obtained from their time lines to create a large document. In our first approach, we simply extracted the document representations of the labeled dataset based on occurrence of 5000 terms as features. Then a Naive Bayes model is used for classification of these documents. This results in some high dimensional and sparse representation for documents.

Table 1: The performance results for the two employed document representations

	Precision	Recall	F1-Measure
Bag of Words	0.69	0.79	0.74
Word2Vec	0.76	0.82	0.79

In the second approach we use a more succinct representation for documents. In this approach first we train word2vec model over our set of unlabeled user time lines. As described before, this is an unsupervised model that can provide some vectorized representation for each word which is also semantically meaningful. Having this word representations we build a proper representation for documents in our labeled dataset.

Different approaches could be considered for building the document representations based on the vectors obtained from word2vec. For example the simplest one is using the bag of words representation and replacing the ones with word2vec vectors and zeros with a zero vectors with a length equal to word2vec vectors. However this method results in very lengthy document vectors. However, considering the few amount of labeled data, this leads to low classification performance. There are other suggested approaches such as using average of the word vectors in a document. We applied this approach to create the feature vector representing each document. Finally, like the bag of words, we use the Naive Bayes model for classifying the extracted document representations.

3 Implementation and Results

We used *Python* along with *sklearn* package as our main machine learning package beside *gensim* package which contains an implementation of word2vec in Python. We conducted two experiments in this work. As our first experiment we used the bag of words model where in the second experiment we used the word vectors obtained from the a pre-training phase by using word2vec method.

For the bag of word model, we represented each document by 5000 terms as features. We used 80 percent of randomly selected labeled dataset for training and its remaining as the test set. We applied a Naive Bayes model over the training set for classification. The results are represented in Table 1.

In the next experiment we used the document representation based on word2vec in our classification. We used word2vec for pretraining and extracting word vectors from a set of timelines of 7237 twitter users. Using the average vectors as described in the previous section we ended up with representing each document by a vector of length 200. Again we used 80 percent of labeled data for training and the remaining for testing. RandomForest model is used for classification. The results are represented in table 1.

As shown in table 1 pre training and proposed document representation

improved all of performance measures. Beside these improvements we may also note that by using pre training we achieved the higher performance by using a highly dense vector representation of only 200 features.

4 Conclusion and Future Works

Here we presented our work on detecting twitter users jobs based on their tweets. This is a new application of tweeter data. Due to the large number of job titles and job categories, in this preliminary work we tried to just recognize people who have an IT related job title. By using the auto labeling method that we described in the data gathering section, we can add more labeled data for other jobs to recognize other job categories as well.

The other contribution in this work is using deep learning to produce document representations and showing that it can produce better results than bag of words model for this specific application. We used word2vec to extract word vectors and proposed a new document representation based on these vectors. We have adopted rather a naive approach for combining the word vectors to create a document representation which can be improved. We could concatenate the word vectors then using dimension reduction techniques we create a fixed length documents that can be fed into classifiers.

We have started to extend the job categories by extracting more data and labeling them by using our cross checking mechanism between Linkedin and Twitter. We also extend the job title dictionary by including the job titles from Canada NoC. This includes about 40000 job categories and job titles.

Another path for extending this work which we are pursuing is using Convolution Neural Networks over the word vectors word2vec. Based on [3, 4] we think that using this approach may produce superior results.

References

- [1] word2vec, <https://code.google.com/p/word2vec/>.
- [2] AMIRI, H., AND III, H. D. Target-dependent churn classification in microblogs. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA*. (2015), pp. 2361–2367.
- [3] KALCHBRENNER, N., GREFENSTETTE, E., AND BLUNSOM, P. A convolutional neural network for modelling sentences.
- [4] KIM, Y. Convolutional neural networks for sentence classification. In *EMNLP’14* (2014), pp. 1746–1751.
- [5] KOULOUMPIIS, E., WILSON, T., AND MOORE, J. Twitter sentiment analysis: The good the bad and the omg! In *Proceedings of the Fifth International*

- Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain, July 17-21, 2011* (2011), AAAI Press, pp. 538–541.
- [6] LI, C., AND SUN, A. Fine-grained location extraction from tweets with temporal awareness. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval* (New York, NY, USA, 2014), SIGIR '14, ACM, pp. 43–52.
 - [7] MIKOLOV, T., TAU YIH, W., AND ZWEIG, G. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT-2013)* (May 2013), Association for Computational Linguistics.
 - [8] SPINA, D., GONZALO, J., AND AMIGÓ, E. Learning similarity functions for topic detection in online reputation monitoring. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval* (New York, NY, USA, 2014), SIGIR '14, ACM, pp. 527–536.